基于互信息和邻接熵的新词发现算法 *

刘伟童 [1, 2]，刘培玉 [1, 2†]，刘文锋 [1, 3]，李娜娜 [1, 2]

(1. 山东师范大学 信息科学与工程学院，济南 250358; 2. 山东省分布式计算机软件新技术重点实验室，济南 250358; 3. 菏泽学院 计算机学院，菏泽 274015)

摘 要：如何快速高效地识别新词是自然语言处理中一项非常重要的任务，针对当前新词发现存在的问题，提出了一种从左至右逐字在未切词的微博语料中发现新词的算法。通过计算候选词语与其右邻接字的互信息来逐字扩展，得到候选新词；并通过计算邻接熵、删除候选新词的首尾停用词和过滤旧词语等方法来过滤候选新词，最终得到新词集。解决了因切词错误导致部分新词无法识别以及通过 n-gram 方法导致大量重复词串和垃圾词串识别为新词的问题，最后通过实验验证了该算法的有效性。

关键词：新词发现；互信息；邻接熵；微博语料

中图分类号：TP301.6 doi: 10.3969/j.issn.1001-3695.2017.11.0745

New word discovery algorithm based on mutual information and branch entropy

Liu Weitong[1, 2], Liu Peiyu[1, 2†], Liu Wenfeng[1, 3], Li Nana[1, 2]

(1. *School of Information Science & Engineering Shandong Normal University*, *Jinan* 250358, *China*; 2. *Shandong Provincial Key Laboratory for Distributed Computer Software Novel Technology*, *Jinan* 250358, *China*; 3. *School of Computer Science*, *Heze University*, *Heze Shangdong* 274015, *China*)

Abstract: How to identify new words quickly and efficiently is a very important task in natural language processing. Aiming at the problems existing in the discovery of new words, there is an algorithm for word-finding new words verbatim from left to right in the uncut word Weibo corpus. One way to get a candidate new word is by computing the candidate word and its right adjacent word mutual information to expand word by word; There are some ways to filter candidate new words to get new word sets. The included methods include calculating the branch entropy, deleting stop words contained in the first or last word of each candidate new word and deleting old words included in the candidate new word set. It solves the problem that some new words can not be recognized due to the mistakes in the word segmentation and It also solves the problem that the large number of repetitive word strings and rubbish words strings generated by the n-gram method are identified as new words. Finally, experiments verified the effectiveness of the algorithm.

Key words: new word discovery; mutual information; branch entropy; microblog corpus

0 引言

随着科学技术的迅速发展，人们通过微博来发表个人意见的情况也越来越常见，大多数的微博都比较随意，非常口语化且不正规，在这种情况下就会产生许多的网络新词。如“屌丝”、“给力”、“尼玛”等。在自然语言处理中，新词的出现对于情感词典的构建、短文本的倾向性分析、中文分词等诸多方面带来了许多不利的影响，降低了它们的效能。所以如何高效的识别新词成为自然语言处理过程中一项非常重要的任务。

目前新词并没有准确的定义，在本文中将未登录词[1]与新词等同，也就是说在本文中新词就是指不在旧词典中的词语。本文使用的旧词典为第六届中文倾向分析评测(COAE) 任务 3 中公开的旧词典资源。

当前新词发现方法共有三种:基于规则的新词发现方法/基于统计的新词发现方法和基于规则与统计相结合的新词发现方法。基于规则的新词发现方法[2,3]是指利用词性特征、语言学的构词规则等方面发现新词，新词发现的准确率较高，但可扩展性、灵活性都比较差，而且还会消耗大量的人力和物力。基于

收稿日期：2017-11-20；修回日期：2018-01-10 基金项目：国家自然科学基金资助项目（61373148，61502151）；山东省社科规划项目（17CHLJ18，17CHLJ33，17CHLJ30）；山东省自然科学基金资助项目（ZR2014FL010）；山东省教育厅基金资助项目（J15LN34）

作者简介：刘伟童（1993-），女，山东潍坊人，硕士研究生，主要研究方向为情感倾向性分析；刘培玉（1960-），男（通信作者），山东潍坊人，教授，博导，主要研究方向为自然语言处理与网络信息安全（liupy@sdnu.edu.cn）；刘文锋（1978-），男，讲师，博士研究生，主要研究方向为自然语言处理；李娜娜（1991-），女，山东济宁人，硕士研究生，主要研究方向为文本摘要提取.

chinaXiv:201804.02058v1

chinaXiv:201804.02058v1

统计的新词发现方法[4~7]是指通过大量的实验语料计算词语的词频、成词概率、左右邻接熵、邻接变化数等统计特征来识别新词。基于统计的方法较为灵活，不受领域的限制、易扩展且可移植性较好，但存在数据稀疏和准确率较低的缺点。基于规则与统计相结合的新词发现方法[8-11]则是希望融合上述两种方法的优点，从而提高新词发现的准确率和效率。

当前主流的新词发现方法是基于规则与统计相结合的方法，充分利用规则和统计这两种方法的优点，期望可以使准确率和效率达到最优。郑家恒等人[2]提出了基于构词法进行新词识别的方法，通过汉语构词的方法建立规则库来获得新词。陈飞等[4]提出的基于条件随机场方法的开放领域新词发现方法，首先用切词工具进行分词，得到标有词性的词语，之后再计算词语的特征值，并利用 CRF 进行学习预测，最终得到新词。该方法仅仅是通过统计的方法来进行新词发现，并没有结合语言规则，而且比较依赖切词系统，若切词系统无法正确识别词语，就会降低新词发现的效果。李文坤等人[5]提出的基于词内部结合度和边界自由度的新词发现方法，首先利用 NLPIR 汉语分词系统[12]对实验语料进行切词处理，计算两个相邻散串的互信息值，之后再利用左右邻接熵进行扩展、过滤，最终得到新词集。夭荣朋等人[8]提出的基于改进互信息和邻接熵的微博新词发现方法，首先通过 n-gram 对语料进行切分得到候选词语，利用词语的词频和停用词等规则进行过滤，之后再利用改进的互信息对候选词语进行扩展，通过计算左右邻接熵值对候选词语进行二次过滤，最后过滤掉旧词语得到新词。该算法利用了规则与统计相结合的方法，但是通过 n-gram 会产生大量的候选词串，导致新词发现的过程比较慢，且会产生较多的垃圾词串，使其准确率比较低。周超等人[10]提出的融合词频特性及邻接变化数的微博新词识别方法，首先利用中科院的切词工具 ICTCLAS 对微博语料进行切词，之后找出两个停用词之间的词串，求得相邻的词串的词频，根据阈值删选得到候选词串，再利用成词规则进行筛选得到候选新词，最后通过邻接变化数过滤，最终得到新词集，该算法比较依赖切词系统，会因为切词错误导致部分新词无法识别。雷一鸣等人[11]提出的面向网络语言基于微博语料的新词发现方法，首先通过切词工具对语料进行切分，获得切词后的散串，并对散串进行统计分析，通过计算散串间的互信息值获得候选新词，并通过过滤低频词和获得邻接变化数的值，进行过滤最终得到新词，该算法也比较依赖切词系统，会因切词系统的好坏决定新词识别的效果。

综上所述，若采用常用的切词系统对实验语料进行切词，有可能会出现因为错误的切分导致新词无法正确识别的现象，从而降低了新词发现的准确率；若采用 n-gram 方法对实验语料进行切分则会出现大量的重复词串，导致新词发现的准确率、效率比较低。所以本文采用在未切词的实验语料中从左到右进行逐字扩展，利用互信息、左右邻接熵等统计特征和删除候选新词的首尾停用词、过滤旧词语等规则相结合进行新词发现。

1 主要技术

1.1 互信息

词语作为一个可以独立存在的语言单元，词语的各个字之间存在一定的相关性。所以字与字或词与字之间的相关性越大，说明字与字或词与字成词的概率也就越大。互信息可以计算两个物体相互依赖的程度，而且互信息值越大，代表两个物体的依赖程度也就越大，所以可以用互信息计算新词的内部成词概率。互信息[13]的计算公式为：

$$MI(x,y)=\log\frac{p(x,y)}{p(x)p(y)} \tag{1}$$

其中:$p(x)$、$p(y)$表示字或词 x、y 单独出现在语料集中的概率，$p(x,y)$表示 x 和 y 共同在语料集中出现的概率。$MI(x,y)$表示 x 和 y 的关联程度。若 $MI(x,y)>0$,表示 x 和 y 是相互关联的，而且 MI 的值越大表示二者相关联的程度越大，也就越有可能成为新词；若 $MI(x,y)=0$，则表示 x 和 y 是彼此独立的；若 $MI(x,y)<0$,则表示 x 和 y 是不相关的。

互信息可以用来计算两个事物的关联程度，所以互信息可以用于发现二元新词，但却无法处理三元及以上的新词。文献[14]经过大量的语料进行新词发现的实验后归纳出 11 种构词模式，而且发现单字模式“1+1”、“1+1+1”、 “1+1+1+1”占新词总数的 61.4%，模式“2+1”、“3+1”占新词总数的 31.2%。从上述可以看出三元及以上的新词在新词的总数中占有一定的比例，如何识别三元及以上的新词是一个亟待解决的问题。

为了解决上述问题，提出了从左至右在未切词的实验语料中逐字扩展的方法。具体方法如下：假设当前字的位置为 i（i=1,…,n），假设相邻字 c_i 和 c_{i+1} 的互信息值 $MI(c_i, c_{i+1})$大于阈值 MI_TH，则继续向右扩展，计算 $MI(c_i\&c_{i+1}, c_{i+2})$，若 $MI(c_i\&c_{i+1}, c_{i+2})$也大于 MI_TH，再继续向右扩展，当且仅当 $MI(c_i\&c_{i+1}\&c_{i+2}\&\ldots\&c_{i+m}, c_{i+m+1})$小于阈值 MI_TH，则停止扩展，并将候选词语 $Can(c_i\&c_{i+1}\&c_{i+2}\&\ldots\&c_{i+m})$加入到候选新词集。

上述方法解决了互信息仅能统计两个元素的局限性，使其对于多元词语也可以进行很好的判断。如：以新词“细思恐极”举例，当统计出候选二元新词 “细思”时，向右扩展，计算“细思”与“恐”的互信息，若高于阈值，则继续向右扩展，计算“细思恐”与“极”的互信息，以此得出新词“细思恐极”，而且上述方法可以避免垃圾词串 “细思”、“细思极”、“恐极”、“思恐极”的产生，极大地提高了新词发现的效率。

1.2 邻接熵

当前确定新词左右边界的方法一般有两种，邻接熵(Branch Entropy，BE)和邻接变化数(Accessor Variety，AV)，本文采用左右邻接熵来确定新词的左右边界。邻接熵[15]可以衡量候选新词的左右邻接字符的不确定性，其不确定性越大，说明其邻接字符包含的信息越多，其成词的概率就越高。

左邻接熵：

$$H_L(W) = -\sum_{W_l \in S_l} P(W_l \mid W)\log p(W_l \mid W) \tag{2}$$

右邻接熵：

$$H_R(W) = -\sum_{W_r \in S_r} p(W_r \mid W)\log p(W_r \mid W) \tag{3}$$

其中：s_l 是候选词 W 的左邻接字的集合，s_r 是候选词 W 的右邻接字的集合； $p(W_l \mid W)$ 表示 W_l 是候选词 W 的左邻接字的条件概率， $p(W_r \mid W)$ 表示 W_r 为候选词 W 的右邻接字的条件概率。其中 $p(W_l \mid W)$ 和 $p(W_r \mid W)$ 的计算公式为

$$p(W_l \mid W) = \frac{N(W_l, W)}{N(W)}, p(W_r \mid W) = \frac{N(W, W_r)}{N(W)} \tag{4}$$

其中： $N(W_l, W)$ 表示 W_l 和 W 共同出现的次数， $N(W)$ 表示 W 出现的次数。同理， $N(W_r, W)$ 表示 W 和 W_r 共同出现的次数， $N(W)$ 表示 W 出现的次数。

本文通过邻接熵来过滤候选新词，但并不用于扩展候选新词，也就是说若候选新词的左邻接熵和右邻接熵都大于阈值，则保留候选新词，否则便删除候选新词。

2　基于互信息和邻接熵的新词发现算法

2.1　本文的改进思路

切词系统日益成熟，但仍不可避免的存在切词错误的现象，在切词后的语料上进行新词发现可能会出现因错误切词导致新词无法识别的现象，可以通过采用 n-gram 方法对语料进行切分，规避因切词系统的错分导致部分新词无法识别的现象。但采用 n-gram 方法对语料进行切分，会出现大量的重复词串，而且将 n-gram 与互信息结合虽然可以用于发现二元新词，且效果极佳，但却无法识别三元及以上的新词。

为了解决上述问题，本文采用逐字从左向右在未切词的语料集上计算互信息的方法来识别新词。解决切词系统、n-gram 方法和互信息存在的缺点。具体步骤如下：

首先是微博语料的预处理。将实验语料按标点符号、特殊符号和英文字母进行切分，并去掉标点符号、特殊符号及英文字母。

其次是计算互信息值，生成候选新词集。从左至右逐字扩展候选词语，因为互信息存在无法有效识别低频词的缺点，所以先统计候选词语与右邻接字共现的词频，按照词频阈值进行过滤，之后再计算候选词语与右邻接字的互信息值，若大于互信息阈值便继续向右扩展，否则就将候选词语记作候选新词。

最后是过滤候选新词，得到新词集。用邻接熵进行候选新词的过滤，设置左右邻接熵的阈值，计算左右邻接熵，将左右邻接熵值小于左右邻接熵阈值的候选新词删除；之后再删除剩余候选新词首尾出现的停用词并过滤掉包含数字的候选新词；因为是发现新词，那么新词必不在旧词典中，所以再过滤掉候选新词集中的旧词语。最后是过滤掉候选新词集中长度小于二的词语，这样便得到了新词集。

2.2　算法流程

本文提出的新词发现算法是面向微博的，因微博自身的特点，微博语料中存在许多的冗余数据，所以首先需要对语料进行预处理，删除微博语料中的冗余成分，并将微博切分成多个短句；然后在预处理后的语料中从左至右逐字计算词与右邻接字的互信息，得到候选新词，最后再利用邻接熵、过滤候选新词首尾的停用词、过滤旧词语等规则进行筛选，最终得到新词集。具体算法的流程如图 1 所示。

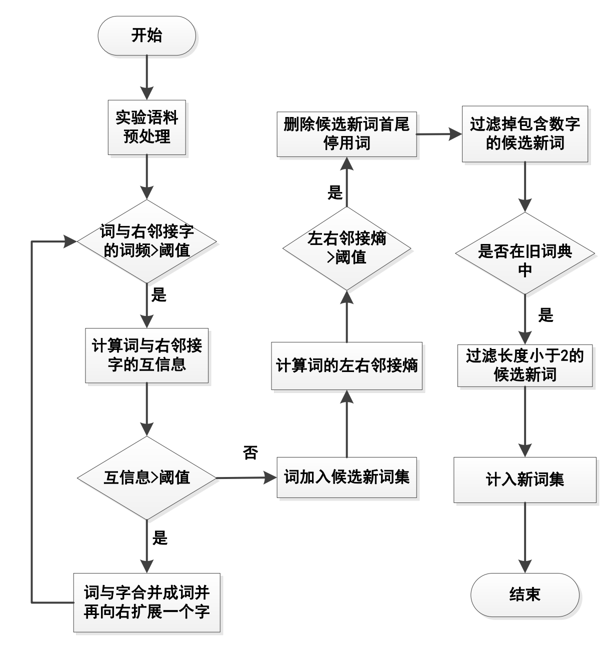


图 1　算法流程图

2.3　算法实现

1）预处理的过程

因为微博中含有大量的噪声数据，所以首先是进行预处理。将微博按照标点符号、特殊符号及英文字母进行断句，将每条微博断成多条小短句，并去掉用于断句的标点符号、特殊符号和英文字母。

输入：微博语料 M

输出：预处理后的微博语料 M_1

M_1=RemoveSymbol(M);　　/*按照 M 中的标点符号及特殊符号断句，并删除符号*/

M_1=RemoveAlpha(M_1);　　/*按照 M_1 中的英文字母断句，并删除英文字母*/

Return　M_1;　　//返回预处理后的微博语料 M_1

2）通过互信息生成候选新词的过程

将微博语料从左至右逐字扩展，统计词语和右邻接字共现的词频，若高于词频阈值再计算候选词语与右邻接字的互信息，若大于互信息阈值便将候选词语与右邻接字组成新的候选词语，并继续向右扩展统计词频并计算互信息值，直到候选词语与右邻接字的互信息小于互信息阈值，则停止向右扩展，记该候选词语为候选新词，从而形成候选新词集。

输入：预处理后的微博语料 M_1

chinaXiv:201804.02058v1

ChinaXiv合作期刊
chinaXiv:201804.02058v1

```
输出：候选新词集 CanList
CanWord=d[0];    /*设置 Canword 的初值为 M1 中某个句子的第一个字*/
for (int  i=1; i<M1.length; i++)
{  //d[i]表示 M1 中某个句子的第 i+1 个字
  CanWord+=d[i];
  if(Freque( CanWord, d[i+1]) >Fre_TH )
  {
    MI(CanWord, d[i+1]);  //计算互信息值
    if(MI( CanWord, d[i+1])<MI_TH)
    {  //MI_TH 为词频阈值
      CanList.add(CanWord);
      Canword=d[i+1];//开始生成下一个候选新词
    }
  }
  else
  {
    Canword=d[i+1];
  }
}
return  CanList;  //返回候选新词集 CanList
```

3)通过邻接熵和部分规则进行过滤得到新词集的过程

对候选新词集按照邻接熵和部分规则进行过滤，设置左右邻接熵的阈值，将大于左右邻接熵阈值的候选新词保留，再删掉候选新词首尾出现的停用词，过滤掉包含数字的候选新词，再过滤掉候选新词中包含在旧词典中的词语，最后过滤掉长度小于二的候选新词，最终得到新词集。

```
输入：候选新词集 CanList
输出：新词集 NewWordList
for (W: CanList)
{  // 对于 CanList 中的每个候选新词计算左右邻接熵
```

$$H_L(W) = -\sum_{W_l \in S_l} P(W_l \mid W)\log p(W_l \mid W)\ ;$$

$$H_R(W) = -\sum_{W_r \in S_r} P(W_r \mid W)\log p(W_r \mid W)\ ;$$

```
  if ( H_L(W) >HL_TH && H_R(W) >HR_TH)
  {  //HL_TH 和 HR_TH 为左右邻接熵阈值
    CanList1.add(w) ;
  }
}
for (W: CanList1)
{  //删除候选新词首尾的停用词
  W=RemoveStopWord(W);
}
for (W: CanList1)
{  //过滤掉包含数字的候选新词
  if( ! W.contain(数字))
    CanList2.add(W);
}
for (W: CanList2)
{  //过滤掉旧词语
  if( ! OldWordList.contain(W))
    CanList3.add(W);
}
for (W: CanList3)
{  //过滤掉长度小于二的词语
  if(W.length>1)
    NewWordList.add(W);
}
return  NewWordList;  //返回新词集 NewWordList
```

3 实验

由于识别网络新词所用的语料并没有比较权威的语料，所以本文实验所用的微博语料是通过爬虫工具采集的 2017 年 3 月到 9 月的新浪微博中比较活跃的部分微博用户（此处将比较活跃的用户定义为每周内至少发六条微博的用户）发表的 10 万条微博。本文采用的算法评价指标有准确率 P(Precision)、召回率 R(recall)和 F 值(F-measure)。

$$P = \frac{N \cap M}{N} \times 100\% \tag{5}$$

$$R = \frac{N \cap M}{M} \times 100\% \tag{6}$$

$$F = \frac{2PR}{P + R} \tag{7}$$

其中:N 表示实验获得的新词的总个数，M 表示微博语料中存在的新词的总个数。

为了证明本文算法的有效性添加了两个对比实验，第一个实验是通过 n-gram 对微博语料进行切分，从而得到候选词，通过计算相邻候选词的互信息来得到候选新词，之后通过计算邻接熵过滤候选新词、删除候选新词首尾出现的停用词、过滤旧词语等获得最终的新词集。第二个实验是文献[11]所用的方法，通过中科院的 ICTCLAS 对微博语料进行切词，并计算互信息得到候选新词，之后通过邻接变化数对候选新词进行过滤，最后过滤掉旧词语，得到新词集。实验结果如表 1，图 2、3 的柱状图更加直观地展示了实验结果。

表 1 实验结果

实验方法	准确率	召回率	F 值	时间
n-gram+MI+BE	55.45%	69.61%	61.73%	2391 s/千条
文献[11]	78.10%	55.87%	65.14%	707 s/千条
本文方法	85.53%	60.78%	71.06%	1021 s/千条

由表 1 和图 2 可以看出，n-gram + MI +BE 方法的召回率

chinaXiv:201804.02058v1

是最高的，即可以正确识别的新词数量是最多的，但准确率的效果却不是很理想，最终导致 *F* 值效果也不佳。主要原因是使用 n-gram 方法对微博语料进行切分，产生了大量的部分重复的词串，这样在最终得到的新词集中包含的词语也就最多，而本文方法也存在一定的缺陷，使得 n-gram + MI +BE 可以识别的正确的新词数比本文方法多，导致召回率高于本文方法。当然，不可避免的在新词集中也会存在许多的垃圾词串，导致准确率效果远低于本文方法的准确率。最终导致本文方法的 *F* 值的效果比 n-gram + MI + BE 的效果好。由表 1 可以看出，n-gram + MI + BE 方法的运行速度是最慢的。综合各方面来说，本文方法优于 n-gram + MI + BE 的方法。

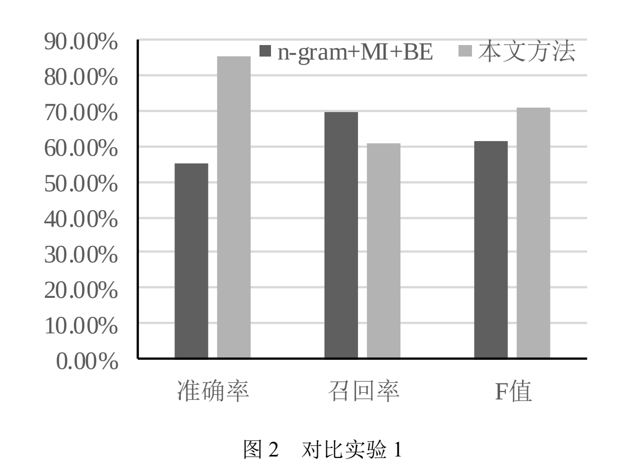


图 2　对比实验 1

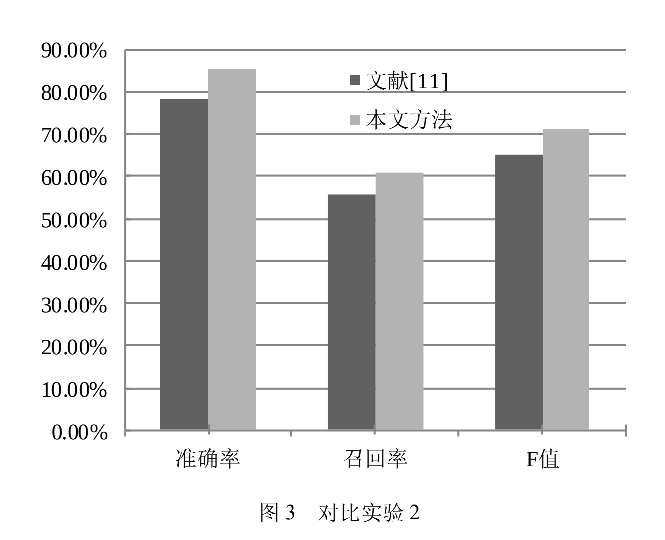


图 3　对比实验 2

通过表 1 和图 3 可以看出，文献[11]的运行速度是最快的，主要原因是切词系统将微博语料切分成许多词串，所以在识别新词的实验中，运行速度会高于本文方法。但文献[11]的方法的准确率、召回率和 *F* 值都不如本文的方法，主要原因是切词系统的错误切分，导致部分新词无法正确识别，使准确率和召回率的效果都不如本文的方法，整体来说，本文方法新词发现的效果优于文献[11]。

由表 1、图 2 和 3 可以看出，虽然本文算法的实验结果的召回率和运行速度不是最佳的，但本文方法的准确率和 *F* 值却是最好的。本文算法采用从左至右逐字来扩展获得新词的方法，可以避免像 n-gram 那样切分数据产生大量候选词串的现象，极大地避免了垃圾词串的产生并提高了运行速度；而且本文不采用任何的切词系统，避免了因切词错误导致部分新词无法识别的现象。从整体来看，本文算法优于上述两种算法，取得了不错的效果。

4　结束语

本文采用从左至右在未切词的微博语料中进行逐字扩展并计算互信息的方法，避免了因通过 n-gram 对语料进行切分时，产生大量重复、无用词串，导致准确率较低和运行速度较慢的现象，同样也避免了因切词系统错误分词导致新词无法识别的现象。通过计算邻接熵、过滤候选新词首尾停用词和旧词语等方法来过滤候选新词，提高新词发现的准确率、召回率和 *F* 值，通过实验也验证了本文算法的有效性。当然本文方法也存在不足，并不能非常准确的识别低频词，由于本文未对微博语料切词，所以在起初判断为词语时，无论是旧词还是新词，都是采用相同的方法，但旧词的词频都相对较高，低频词相对较少，所以本文方法发现旧词会比发现新词的能力强，但因低频词的影响，未能识别所有的旧词。所以本文的新词发现算法还存在可提高的空间，希望以后可以针对低频词提出改进意见，再进一步的调高算法的准确率。

参考文献：

[1] Chen Kehjiann, Ma Weiyun. Unknown word extraction for Chinese documents [C]// Proc of the 19th International Conference on Computational Linguistics. 2002: 1-7.

[2] 郑家恒，李文花．基于构词法的网络新词自动识别初探 [J]. 山西大学学报：自然科学版, 2002, 25 (2): 115-119.

[3] 李明．针对特定领域的中文新词发现技术研究 [D]. 南京：南京航空航天大学, 2012.

[4] 陈飞，刘奕群，魏超，等．基于条件随机场方法的开放领域新词发现 [J]. 软件学报, 2013, 24 (5): 1051-1060.

[5] 李文坤，张仰森，陈若愚．基于词内部结合度和边界自由度的新词发现 [J]. 计算机应用研究, 2015, 32 (8): 2302-2304.

[6] 霍帅，张敏，刘奕群，等．基于微博内容的新词发现方法 [J]. 模式识别与人工智能, 2014, 27 (2): 141-145.

[7] Peng F, Feng F, Mccallum A. Chinese segmentation and new word detection using conditional random fields [J]. Proceedings of Coling, 2004: 562-568.

[8] 夭荣朋，许国艳，宋健．基于改进互信息和邻接熵的微博新词发现方法 [J]. 计算机应用, 2016, 36 (10): 2772-2776.

[9] 林自芳，蒋秀凤．基于词内部模式的新词识别 [J]. 计算机与现代化, 2010 (11): 162-164.

[10] 周超，严馨，余正涛，等．融合词频特性及邻接变化数的微博新词识别 [J]. 山东大学学报：理学版, 2015, 50 (3): 6-10.

[11] 雷一鸣，刘勇，霍华．面向网络语言基于微博语料的新词发现方法 [J]. 计算机工程与设计, 2017, 38 (3): 789-794.

ChinaXiv合作期刊

[12] 张华平. NLPIR 汉语分词系统 [EB/OL]. http://ictclas. nlpir. org/.

[13] Ye, Yunming, Chow, *et al.* Unknown Chinese word extraction based on variety of overlapping strings [J]. Information Processing & Management, 2013, 49 (2): 497-512.

[14] 崔世起. 中文新词检测与分析 [D]. 北京：中国科学院计算技术研究所, 2006.

[15] Huang J H, Powers D. Chinese word segmentation based on contextual entropy [J]. Diseases of the Colon & Rectum, 2000, 4 (6): 402.

chinaXiv:201804.02058v1